Research Report

Confidence and Cognitive Test Performance

Lazar Stankov

Jihyun Lee

Research & Development

March 2007
RR-07-03

Confidence and Cognitive Test Performance

Lazar Stankov and Jihyun Lee
ETS, Princeton, NJ

March 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper examines the nature of confidence in relation to cognitive abilities, personality traits, and metacognition. Confidence was measured as it was expressed in answers to each test item during the administration of reading and listening sections of the TOEFL® iBT. The confidence scores were correlated with the accuracy scores from the TOEFL iBT, SAT®, high school grade point averages (HS-GPA), and several measures of personality and metacognition. The results indicate that confidence is a separate psychological trait, somewhere between cognitive ability and personality traits. In addition, our findings suggest that confidence is related to, but separate from, metacognition. We also demonstrated gender and ethnic differences in confidence, with male and African American students showing higher overconfidence bias than females and White or Hispanic students, respectively. Finally, our data show small but statistically significant incremental validity of the confidence scores above and beyond the accuracy scores in predicting numeracy test scores, the total TOEFL iBT scores, and TOEFL iBT subscores on the writing and speaking sections. We found no incremental validity of the confidence scores in predicting the SAT and HS-GPA. Theoretical and practical implications of these findings are discussed as well.

Key words: Confidence, overconfidence bias, metacognition, self-monitoring

Acknowledgments

We are grateful to Jennifer Minsky, who was involved in many stages of what became known as the New Constructs Study, which provided the data for this paper. We are also grateful to Cathy Trapani and her team of data analysts for data cleaning and the statistical analyses reported here. Useful comments were provided by Henry Braun, Brent Bridgeman, Walter Emmerich, Gerry Fogarty, Sabina Kleitman, Pat Kyllonen, Ida Lawrence, Don Powers, and Larry Stricker at the Turnbull Cluster Brown Bag Seminar at ETS and were incorporated in the current version of this paper.

Table of Contents

List of Tables

Page

Introduction

Psychologists in the field of decision-making and education inspired the work on confidence by questioning whether those who know more also know more about how much they know. Knowing refers to *accuracy* and knowing how much they know relates to *confidence* (Lichtenstein & Fischoff, 1977). Two important theoretical approaches have been dominant in the study of confidence: the heuristics and biases approach (Kahneman, Slovic, & Tversky, 1982) and the ecological approach (Gigerenzer, Hoffrage, & Kleinbolting, 1991). In educational research, the same question is asked under the rubric of metacognition (Schraw & Dennison, 1994; Tobias & Everson, 2000, 2002).

In fact, interest in confidence has a long history in psychology. Psychophysical studies of confidence started with the work of Fullerton and Cattell (1892), Trow (1923), and Festinger (1943a, 1943b). These classical psychophysicists routinely assessed confidence along with accuracy and speed and demonstrated that accuracy, speed, and confidence are highly related in threshold performance. Later, Vickers (1979) and Baranski and Petrusic (1999) supported this early work in psychophysics. However, there has been evidence showing a relative independence of confidence and speed measures in completing complex cognitive tasks (see Stankov, 2000, for the review).

Individual Differences in Confidence[1]

Most of earlier work on confidence has been experimental rather than differential. Only fairly recently (i.e., the late 1990s), the work on decision-making has addressed individual differences (see Soll, 1996, for the review). A good number of studies followed, showing pronounced individual differences in confidence ratings (Pallier et al., 2002; Schraw & Dennison, 1994; Stankov, 1998, 1999; Stankov & Crawford, 1996, 1997; Stanovich, 1999). Within this framework, confidence is considered as a disposition, in other words, a systematic tendency that leads one to act in a particular way because of his or her belief in oneself. Those scoring high on confidence measures are described as decisive, firm, and resolute while those scoring low are described as indecisive, doubtful, and vacillating about their decisions and capacity. On the other hand, Stankov (1999) argued that the essence of confidence cannot be reduced to either disposition or cognition, but rather it can only be placed on no man's land between personality and abilities, along with other constructs such as various self constructs, intellectual engagement, and perhaps emotional intelligence. In addition, Crawford and Stankov

(1996) showed that objectivity of people's self-confidence ratings can be reliably measured if one assesses confidence in typical testing-taking situations and compares its ratings to the actual cognitive performance. In other words, although people use reality checks at the subjective level by asking themselves *"How confident I am that my answer is correct?"* confidence can be assessed with an objective method when it is measured over sets of items in a test. The following section elaborates on this.

The Measurement of Confidence

The procedure employed in the present study for assessing confidence follows Crawford and Stankov's (1996) approach. Participants are asked to give a rating (expressed in terms of percentages) immediately after responding to an item in a test to indicate how confident they are that their chosen answer for this item is correct (see Crawford & Stankov; Harvey, 1997; Keren, 1991; Stankov, 1999). Thus, these ratings directly follow the cognitive function of providing an answer. Confidence ratings for the attempted test items are averaged to give an overall confidence score. Confidence scores have been studied both on their own or in relationship to the accuracy measures obtained from the same cognitive test (Kleitman & Stankov, 2001; Stankov, 2000).

In studying the relationship between confidence and accuracy on a typical cognitive test, accuracy scores are often converted to average percentage correct and subtracted from the average percent confidence scores. The result is called bias score.[2] The considerable amount of data show overconfidence bias in cognitive tasks—that is, pronounced positive bias scores resulting from the difference between confidence and accuracy scores (see Lichtenstein & Fischoff, 1977; Lichtenstein, Fischoff, & Phillips, 1982). Pronounced underconfidence bias has been found with some sensory tasks (Olsson & Winman, 1996). However, we use the term bias interchangeably with overconfidence bias in the present study, since all tasks employed in this study are cognitive which are known to display overconfidence (Stankov, 2000). Zero is often treated as an ideal value for bias scores because a good match between the level of confidence and performance is often seen as a desirable characteristic. In this paper, the unit of analysis is arithmetic means of bias scores (i.e., differences between the means for accuracy and confidence scores) to highlight the discrepancies between subpopulations. We do not employ bias scores at the individual level of analysis (i.e., in analyzing correlations).

Correlates of Confidence

Evidence suggests that confidence shows structural independence from other established ability and personality traits. Also, it has been shown that confidence is a general trait across different tasks (see Blais, Thompson, & Baranski, 2005; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997). Three main domains have been most often related to the construct of confidence. First, confidence scores have shown moderate correlations with measures of *cognitive abilities*—that is, a higher accuracy score is linked to a higher confidence level. The list of cognitive tasks that have been studied in relation to confidence includes measures of verbal and nonverbal reasoning (fluid intelligence, *Gf*), acculturated knowledge (crystallized intelligence, *Gc*), long-term and short-term memory, visual, olfactory, gustatory, and auditory perceptual processes, among others. In general, the studies show that people who are more confident on one cognitive task tend to be more confident across other tasks. Evidence also indicates that correlations between accuracy and confidence scores from the *same* tests tend to be between .40 and .60 (Stankov, 2000), while correlations among confidence scores from *different* cognitive tests have been equal to, or higher than, correlations between accuracy and confidence scores from the same tests. This suggests that a general confidence factor exists that is separate (yet positively related) to factors of intelligence. Confidence as a separate factor has been identified in studies by Kleitman and Stankov (2001), Pallier et al. (2002), Stankov (1998, 1999), and Stankov and Crawford (1996, 1997); see also Stanovich (1999).

Second, confidence is sometimes treated as a *personality trait*, either on its own or as an underlying facet of broader traits (Blais, Thompson, & Baranski, 2005; Pallier et al., 2002). Consistent small correlations (r = .30) have been noted between confidence and the openness factor from the Big Five model (Pallier et al., 2002). However, it is unclear whether this relationship with openness is mediated by cognitive abilities since both confidence and openness correlate with cognitive abilities. Moreover, Blais et al. demonstrated that a broad range of cognitive styles, including the need for cognition and the desire for structure, did not have an effect on confidence. Conceptually, cognitive styles straddle the boundary between personality and metacognition.

Third, confidence judgments are frequently interpreted as an important aspect of *metacognitive* processes. Metacognition refers to the awareness of one's cognitive strengths and

weaknesses and one's learning processes. One component of metacognition, self-monitoring, refers to the awareness of the accuracy of one's answers on typical cognitive tests. Bias score can be used as a direct measure of self-monitoring (see Schraw & Dennison, 1994). Separate metacognitive factor based on questionnaire data was identified in a number of studies. That factor was distinct from the confidence factor based on the procedure used in the present paper and it has shown moderate correlation with confidence (see Kleitman, 2003; Kleitman & Stankov, 2006).

Gender and Ethnic Differences in Confidence and Bias Scores

Gender differences in confidence and bias scores have been examined extensively, but the findings are less than conclusive so far. Studies by Stankov and Crawford (1996, 1997) found no gender differences, but more recent studies by Pallier et al. (2002) and Pallier (2003) showed significant gender differences in confidence, with females exhibiting lower bias scores on cognitive tasks. This paper addresses this gender difference issue.

The evidence for possible ethnic differences in confidence has been explored only indirectly so far. Stankov (in press) reported that on tests of esoteric analogies and vocabulary, foreign students seeking admission to U.S. universities tend to have overconfidence bias scores that are about twice as high as U.S. students. At present, there is no available information on whether different ethnic groups within a country may show pronounced differences in confidence ratings. The present study examines ethnic differences in confidence and bias scores among White, Hispanic, and African American participants.

Evidence for Incremental Validity of Confidence Scores

The evidence for incremental validity of confidence scores is not available at present. In other words, can confidence predict educational outcomes or job performance after taking account of test scores that are typically employed in admission or selection processes? The studies on confidence to date have been limited to the understanding of psychometric properties of the confidence scores themselves (Kleitman & Stankov, 2006). This study focuses on the issue of predictive validity of confidence scores using various aptitude measures as criteria. These include self-reported SAT® scores and high school GPA (HS-GPA), numeracy test scores, TOEFL® scores for the writing and speaking sections. These selected measures are frequently used to predict both school and job performance.

In showing the incremental validity of confidence scores above and beyond their yoked accuracy scores, we predict accuracy scores from cognitive measures that are not used to extract confidence scores. For instance, the question is whether confidence scores obtained from the reading and listening sections in the TOEFL exam will add to the prediction of HS-GPA above and beyond TOEFL reading and listening accuracy scores.[3] This is a stringent requirement since the accuracy scores from the criterion (i.e., HS-GPA) and the predictor measures (i.e., TOEFL reading and listening sections) are capturing the same construct—acculturated knowledge (or crystallized ability, *Gc*). The presence of incremental validity may contribute to the argument for using confidence scores in admission, selection, or training programs.

Goals

The purpose of this study is to examine the nature of confidence exhibited during performance on two sections (reading and listening) of a cognitive test, the TOEFL iBT exam. Three issues will be addressed. First, the study examines psychometric properties of confidence scores that are obtained from TOEFL iBT reading and listening sections—their reliability and validity. Validity of confidence scores is examined by assessing their relationships to the accuracy scores from several cognitive tests and questionnaire measures of metacognition and personality. The criteria are accuracy scores from TOEFL iBT writing and speaking sections, TOEFL iBT total scores, standardized measures of academic performance (SAT and ACT), HS-GPA, and two additional cognitive tests (numeracy and overclaiming). Second, the study explores gender and ethnic differences on accuracy, confidence, and bias scores. Third, the study examines the incremental validity of confidence scores for predicting cognitive criterion measures employed in this study above and beyond accuracy scores based on TOEFL iBT reading and writing sections.

Method

Participants

Participants (N = 824) in this study were recruited from two types of colleges. One group (N = 371) came from nine 2-year community colleges.[4] The other group (N = 453) came from twelve 4-year colleges. All participants were native speakers of English. There were 304 male and 518 female participants. In terms of ethnic composition, the sample consisted of participants who were White (N = 605), African American (N = 112), Hispanic (N = 60), or other (N = 46).

Procedure

Participants were administered the full TOEFL iBT exam in the morning and were asked to attend another session in the afternoon on the same day at the beginning of 2006. The afternoon session started with the repeated test consisting of the selected reading and listening items with confidence ratings attached to each item (see the next section). To allow participants sufficient time to answer confidence ratings, there was no time limit in the afternoon session. In addition, the participants took a battery of 28 instruments. These additional instruments were measures of cognitive ability, personality, metacognition, interests, emotional intelligence, as well as social attitudes, values, and social norms. In this paper, we only report results based on the measures on cognitive ability, personality, and metacognition from the entire battery of 28 measures along with the repeated items on the TOEFL iBT reading and listening sections.

Instruments

TOEFL iBT (Form B)

The most recent version of the TOEFL exam (known as the TOEFL Internet-based test or TOEFL iBT) consists of four sections—speaking, writing, reading, and listening. The test is delivered via the Internet. Detailed analyses of the TOEFL iBT exam based on nonnative speakers of English are provided in ETS's technical report by Sawaki, Stricker, and Oranje (2005). As has been the case with previous versions of TOEFL, the validation process includes the administration to the native speakers of English (see Angelis, Swinton, & Cowell, 1979; Angoff & Sharon, 1971; Johnson, 1977; Stricker, 2002). The present study is based on such a sample of native speakers from community colleges who took the TOEFL iBT exam. There is considerable research showing that TOEFL performance correlates with other cognitive tests (Stricker). The GRE® verbal reasoning score correlates .61 with the TOEFL total score for nonnative speakers of English, and this correlation is .64 for native speakers (Stricker).

The raw scores from each section are converted to scaled scores from 0 to 30. The total TOEFL iBT score is a simple sum of the four subscale scores, ranging from 0 to 120. The description of these four sections follows.

Reading. The reading section has three item-sets, containing a total of 40 items (12, 14, and 14 in Sets 1, 2, and 3, respectively). Thirty-seven items are four-option multiple-choice items. One item in each set is an open-ended item. The work reported in this paper is based on the multiple-choice items in the first two item sets. Thus, reading, version 1 consisted of 11 items

from the first set and reading, version 2 consisted of 13 items from the second set. Figure 1 presents a screen capture of an item from the reading section and confidence ratings associated with the item.

Listening. The listening section has 34 items in six item sets, with two item sets based on conversation and four item sets on lectures on academic topics. Each stimulus is followed by five items with four multiple-choice options. We selected conversation-based listening items (17 items) for the afternoon session to reduce testing time. Listening, version 1 contains 11 items from two different conversations, and listening, version 2 contains 6 items from another conversation.

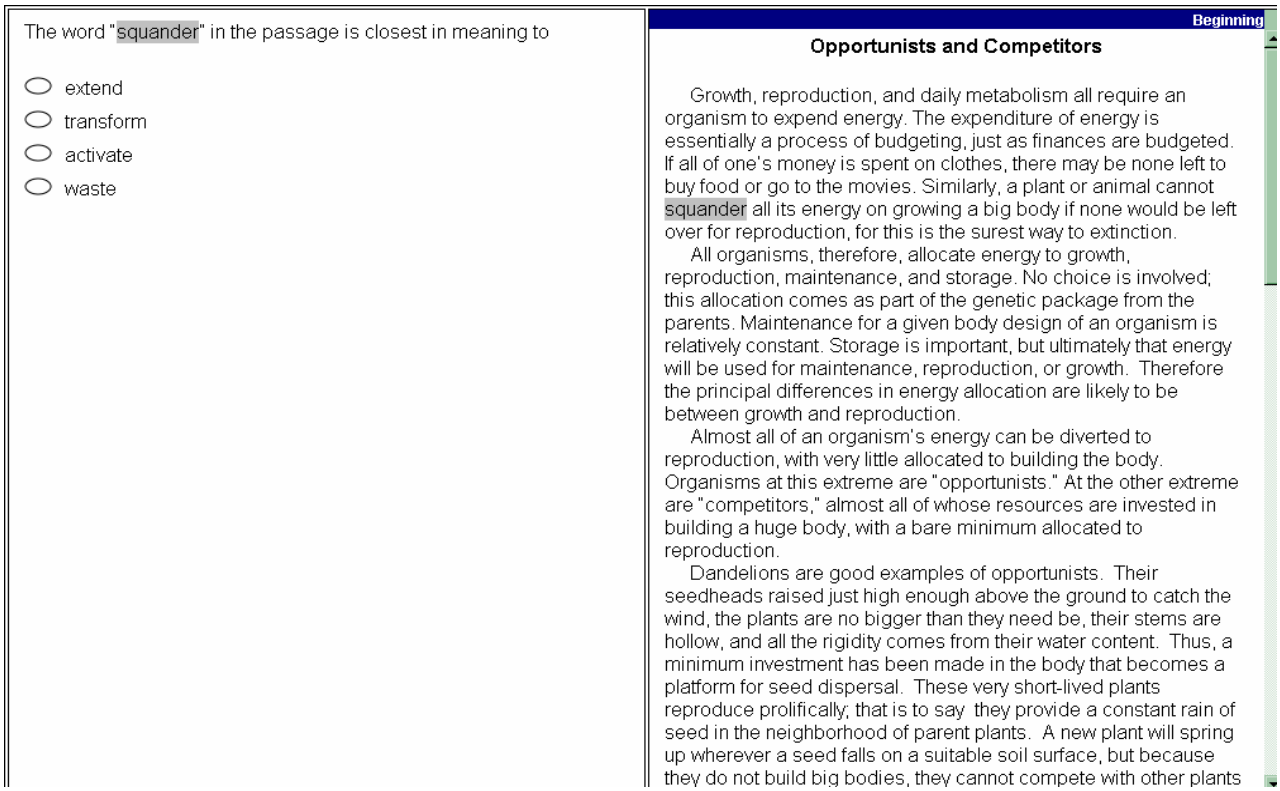

Figure 1. **Screen capture of an item from the TOEFL iBT, Form B.**

Note. After providing the answer to an item, participants are asked to answer confidence question.

Speaking. The speaking section consists of six items. Examinees' responses to each item are scored on a scale of 0–4 by trained raters. The raw speaking score is a sum of the points across all six items, resulting in a score range from 0 to24.

Writing. The writing section includes two items. Examinees' responses to each item are scored on a scale of 0 to 5 by trained raters. The raw writing score is a sum of the points earned on the two items, and thus, the raw writing score ranges from 0 to 10.

Confidence Measures

The confidence scores were obtained during the administration of a TOEFL iBT exam to native speakers of English. Our approach to measuring confidence is to obtain accuracy and confidence scores from the same cognitive test items. We used multiple-choice items from two TOEFL iBT—reading and listening—sections to collect confidence ratings for the present study (see Figure 1_.

Confidence expressed in TOEFL items is certainly related to participants' confidence in their command of English. However, our study is built upon the previous research showing that confidence is in fact a general trait, implying that confidence requires decision-making in all cognitive activities and not just in foreign-language learning situation. Thus, confidence scores based on each item in the TOEFL exam are used for this study, as any other cognitive test can be used.

Additional Ability Measures

Apart from TOEFL iBT, two ability measures used in this study were a numeracy test and an overclaiming test.

Numeracy test. Developed by ETS, *The National Adult Literacy Survey* contains a numeracy test to assess examinees' abilities to use and manipulate numerical information in a real-world context (Kirsch et al., 2001). Seventeen items from this test were used in the present study.

Overclaiming test. This 45-item instrument (Paulhus & Harms, 2004) assesses respondents' tendency to overclaim their familiarity with historical names and events, social sciences, and physical sciences. Each item is on a scale ranging from 0 (never heard of it) to 6 (know it very well). Within each three categories, 3 out of every 15 items are foils, that is, they do not actually exist. Hence, any degree of familiarity with foils constitutes overclaiming. The data

on overclaiming is typically analyzed by the signal detection theory (Paulhus & Harms). Signal detection analysis exploits all of the data in the calculation of separate indexes for *accuracy* and *response bias*. The best known formula for scoring overclaiming is *d-prime*, in which individuals showing the best discrimination about real items relative to foils will get the highest d-prime score as opposed to the ones simply scoring the most hits. We employed this d-prime (*d'*) measure in this study.

Metacognitive Inventories

Two questionnaires were employed to assess metacognition: the *Metacognitive Awareness Inventory (MAI), Memory and the Reasoning Competence Inventory* (MARCI).

Metacognitive Awareness Inventory. Ten items were chosen from the original 52 items of the MAI (Schraw & Dennison, 1994). These items showed satisfactory psychometric properties in a pilot for the present study. The MAI contains questions on students' self-perceptions, their strong and weak points as learners, their learning strategies, and the conditions under which they can learn most effectively.

Memory and Reasoning Competence Inventory. This is a measure of self-concepts of memory and reasoning, consisting of 16 items with 8 items for each component (Kleitman, 2003; Kleitman & Stankov, 2006). The process of scale development is based on a model of self-concept items (Marsh & Shavelson, 1985; see also Marsh, 1986). The respondents evaluate the extent to which each statement describes them using a 6-point Likert scale ranging from false (scaled at 1) to true (scaled at 6). Memory and reasoning items are mixed in order. Separate scores for these two components are calculated following Kleitman's finding of two separate factors in this instrument (i.e., 8 items on memory and 8 items on reasoning). Examples of memory items include: "Compared to other intellectual abilities (i.e., attention, reasoning), my memory is good," and "My memory is above average." Examples of reasoning items include: "I feel confident when solving problems that require reasoning skills," and "I can reason better than the average person."

Personality Measures

For a personality measure, we used the Big Five Personality Inventory scales for extraversion, conscientiousness, agreeableness, emotional stability, and openness, which are available from the International Personality Item Pool (IPIP; n.d.).

Additional Outcome Measures

Participants were also asked to provide their scores on standardized tests (either the SAT or ACT) and their HS-GPA. These are self-reported scores that have not been checked for accuracy.

Results

Descriptive Statistics on the Accuracy and Confidence Scores

Descriptive statistics for accuracy and confidence scores are provided in Table 1. The information on the first presentation in Table 1 was obtained from the multiple-choice items during the full operational TOEFL iBT testing session in the morning. The information about the second presentation was obtained from the same repeated items given in the afternoon session, which included participant-selected confidence ratings for each item. Raw scores were transformed into percentage correct scores by dividing the raw score by the number of test items and multiplying it by 100. The mean column for confidence in Table 1 shows the mean confidence scores across the items in each scale. The bias column in Table 1 presents the difference between the means for confidence and the percentage correct scores.

Table 1

Arithmetic Means for Accuracy, Confidence, and Bias Scores for TOEFL iBT Reading and Listening Sections

Variable	Accuracy			Confidence	Bias
	N	Mean (SD)	% correct	Mean	Mean
1st presentation: TOEFL iBT					
Reading 1	11	8.55 (2.12)	78.78	--	--
Reading 2	13	9.38 (2.58)	72.15	--	--
Listening 1	11	8.84 (2.06)	80.36	--	--
Listening 2	6	4.94 (1.17)	82.33	--	--
2nd presentation: TOEFL iBT with confidence scale					
Reading 1	11	8.78 (2.19)	79.82	88.47	8.65
Reading 2	13	9.44 (2.56)	72.54	87.37	14.83
Listening 1	11	8.98 (2.05)	81.64	87.21	5.57
Listening 2	6	5.01 (1.19)	85.00	90.48	5.48

Note. $N = 824$.

Several observations are noteworthy in Table 1. First, there are remarkably small differences in the mean performance between the morning and afternoon sessions, indicating that the absence of time restriction in the afternoon session did not affect performance to any significant degree, at least at the group-mean level. Second, the reading section was more difficult than the listening, although both were somewhat easy as expected from the sample of native English speakers—the average difficulty levels range between 72% and 85%. Third, the mean variations across four sets of tasks were more salient in the accuracy scores than the confidence scores. The participants' confidence scores were more constant throughout the different tasks. Fourth, reading 2 shows a very high bias score in the present data.[5] Previous studies indicate that bias scores greater than 10% should be considered substatnial (Kleitman & Stankov, 2001; Stankov & Crawford, 1996, 1997).

Reliabilities of the Accuracy and Confidence Scores

Three types of reliability coefficients are presented in Table 2: (a) Chronbach's alphas for the accuracy and confidence scores, (b) parallel form (i.e., correlations between the accuracy scores on versions 1 and 2 for the reading and listening sections), (c) test-retest reliabilities (i.e., correlations between the accuracy scores from the two administrations).

All reliability coefficients for both the accuracy and confidence scores are at an acceptable level for research purposes, except for listening, version 2, which contains only six items. It is noteworthy that the confidence scores show higher reliabilities than the accuracy scores and that the reliabilities of the confidence scores are consistently high across different tasks and different versions.

Effects of Confidence on Changes in Accuracy Scores From Test to Retest

This section explores possible causes of less than perfect correlations in the accuracy scores between the two testing sessions. In particular, we obtained absolute change scores from the differences in the accuracy scores between the morning and afternoon sessions. The absolute change scores indicate the amount of change irrespective of whether there was a reduction or increase in accuracy scores between these two sessions. The arithmetic means for the absolute change scores are shown in Table 3. The correlations between the absolute change and confidence scores are negative with the coefficients in the upper .20s, indicating that more confident people tend to have lower absolute change scores. Thus, people's levels of confidence

seem to affect reliability of the accuracy scores which, in turn, is reflected in test-retest correlations. Changes occurred more frequently on the reading section, which is more difficult, than on the listening section.[6]

Table 2

Reliabilities for TOEFL iBT Reading and Listening Scores With and Without Confidence Scale

Variable	Accuracy				Confidence
	N	Alpha	Parallel	Test-retest	Alpha
1st presentation: TOEFL iBT					
Reading 1	11	.73	--	--	--
Reading 2	13	.79	.70	--	--
Listening 1	11	.75	--	--	--
Listening 2	6	.62	.55	--	--
2nd presentation: TOEFL iBT with confidence scale					
Reading 1	11	.82	--	.85	.91
Reading 2	13	.79	.72	.78	.94
Listening 1	11	.78	--	.71	.94
Listening 2	6	.72	.59	.68	.90

Note. $N = 824$.

Table 3

Pearson Product-Moment Correlations Between Absolute Change Scores Between Test and Retest and Confidence Scores on TOEFL iBT Reading and Listening Sections

Confidence scores	Absolute change scores	
	Reading	Listening
Reading	–.29	–.27
Listening	–.18	–.26
Mean (SD)	2.22 (3.02)	1.58 (2.20)

Factor Analyses of the Confidence Scores

In this section we explore whether the confidence factor will stand separate from cognitive abilities and metacognition. Two analyses were carried out. The first factor analysis employed eight variables—accuracy and confidence scores from four measures (listening, versions 1 and 2 and reading, versions 1 and 2) of the TOEFL iBT exam that were administered in the afternoon session. The second analysis included, in addition to the composite listening and reading scores, two other cognitive ability tests (i.e., numeracy and overclaiming *d'*) and the questionnaire measures of metacognition (MARCI and MAI). The reason for carrying out two separate factor analyses is to avoid the undue influence of parallel forms of the test on the factor structure. In other words, versions 1 & 2 of reading or listening tests are parallel-form estimates and should be summed when both versions are to be used with other measures in factor analysis.

First factor analysis. Table 4 presents correlations among the accuracy and confidence scores from the two versions of the TOEFL iBT reading and listening sections. All correlations are moderately high and positive. Several observations are noteworthy. First, the confidence scores correlate higher among themselves than do the accuracy scores. Second, for both accuracy and confidence scores, the correlations between the two sections (i.e., between versions 1 and 2 of the same test) are higher than the correlations between reading and listening sections. Third, correlations between the accuracy and confidence scores from the same version of the tasks are higher (.445 to .605) than the correlations between accuracy and confidence scores from the different version of the same tasks (.358 to .574). Fourth, there are slightly stronger associations between reading accuracy and confidence scores (.469 to .605) than between listening accuracy and confidence scores (.358 to .490).

Tables 5 and 6 provide a factor pattern matrix on the accuracy and confidence scores, which was obtained by applying the root-one criterion and using maximum likelihood estimation and PROMAX rotation for factor extraction and rotation. It clearly shows that two factors, verbal comprehension and confidence, exist in these data. Factor intercorrelation (.578) appears a bit high, but it is within the expectations especially when we consider the fact that the accuracy and confidence scores are based on the same items.

Table 4

Pearson Product-Moment Correlations Between Accuracy and Confidence Scores, TOEFL iBT

Variable	Accuracy scores				Confidence scores			
	Reading 1	Reading 2	Listening 1	Listening 2	Reading 1	Reading 2	Listening 1	Listening 2
Accuracy scores								
Reading 1	--							
Reading 2	.781	--						
Listening 1	.552	.595	--					
Listening 2	.569	.616	.684	--				
Confidence scores								
Reading 1	.605	.574	.431	.499	--			
Reading 2	.469	.519	.366	.450	.824	--		
Listening 1	.350	.360	.445	.490	.704	.768	--	
Listening 2	.282	.291	.358	.480	.629	.717	.822	--

Table 5

Exploratory Factor Analysis of the Correlations Between Accuracy and Confidence Scores

Variable	Factor	
	Confidence	Verbal comprehension
Accuracy scores		
Reading 1		.918
Reading 2		.967
Listening 1		.598
Listening 2	.215	.570
Confidence scores		
Reading 1	.593	.358
Reading 2	.753	
Listening 1	.962	
Listening 2	.958	

Table 6

Factor Correlation Matrix

Factor	Confidence	Verbal comprehension
Confidence	--	
Verbal comprehension	.578	--

Second factor analysis. Table 7 shows correlations among the accuracy and confidence scores on TOEFL iBT reading and listening sections, versions 1 and 2 combined, and the accuracy scores on the numeracy, overclaiming *d'* tests, and the metacognitive measures (memory inventory, reasoning inventory, and metacognitive awareness inventory). All the correlations in Table 7 are positive and moderate in size except for the ones between metacognitive inventories and overclaiming *d'* scores. As expected, the correlations in Table 7 are lower overall than those presented in Table 4.

Tables 8 and 9 show the factor pattern matrix based on the correlation matrix presented in Table 7. We report maximum likelihood solution followed by PROMAX rotation. Using the root one criterion, three factors were extracted. They are:

1. Acculturated knowledge *(Gc)*. The accuracy scores from all four tests of cognitive abilities load on this factor. Although three of the four tests that load on this first factor involve verbal abilities, loading from the numeracy test indicates that this factor represents crystallized abilities (*gc*). Reading confidence also has a small loading on this factor.

2. Confidence. The confidence scores from the TOEFL iBT reading and listening sections load on this factor. This is a doublet due to the absence of other confidence measures in this study. However, previous work has shown that confidence factor can reliably be extracted from a larger number of different cognitive tasks (see Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997).

3. Metacognition. All three measures of metacognitive processes load on this factor.

The factor correlation between the acculturated knowledge (*Gc*) and confidence factors is moderate in size (r = .552), about the same as the factor correlation reported between verbal comprehension and confidence in Table 5. Metacognitive factor shows lower correlations with both *Gc* and confidence factors (r = .379 and .317, respectively).

Table 7

Pearson Product-Moment Correlations Among Accuracy and Confidence Scores and Metacognitive Inventories

Variable	Accuracy scores				Confidence scores		Metacognitive inventories		
	Reading 1 & 2	Listening 1 & 2	Numeracy	Overclaiming *d'*	Reading 1 & 2	Listening 1 & 2	Memory	Reasoning	Metacognitive awareness
Accuracy scores									
Reading 1 & 2	--								
Listening 1 & 2	.693	--							
Numeracy	.616	.532	--						
Overclaiming *d'*	.375	.304	. 355	--					
Confidence scores									
Reading 1 & 2	.594	.501	.437	.254	--				
Listening 1 & 2	.349	.495	.306	.168	.774	--			
Metacognitive inventories									
Memory	.375	.137	.157	.081	.196	.145	--		
Reasoning	.164	.268	.318	.169	.339	.254	.466	--	
Metacognitive awareness	.322	.019	.028	-.067	.102	.088	.271	.302	--

Table 8

Exploratory Factor Analysis of the Correlations Among Accuracy and Confidence Scores and Metacognitive Inventories

Variable	Factor		
	Acculturated knowledge	Confidence	Metacognition
Accuracy scores			
Reading 1 & 2	.962		
Listening 1 & 2	.640		
Numeracy scores	.630		
Overclaiming *d'* scores	.409		
Confidence scores			
Reading 1 & 2	.314	.618	
Listening 1 & 2		.994	
Inventories			
Memory			.626
Reasoning			.724
Metacognitive awareness			.453

Table 9

Factor Correlation Matrix

Factor	Acculturated knowledge	Confidence	Metacognition
Acculturated knowledge	--		
Confidence	.552	--	
Metacognition	.379	.317	--

Correlations Between TOEFL iBT Reading and Listening Accuracy and Confidence Scores With TOEFL iBT, SAT, ACT, and HS-GPA

Table 10 presents correlations between a composite confidence score (i.e., the sum of the confidence scores from the TOEFL iBT listening and reading sections) and other cognitive performance measures, the total score on TOEFL iBT, self-reported SAT, and ACT scores, and HS-GPA.

As can be expected, correlations of the confidence scores are higher with the scores from TOEFL iBT reading (.499) and listening (.539) sections than with the scores from TOEFL iBT speaking (.338) and writing (.414) sections. The confidence scores correlate moderately with the total TOEFL iBT scores (r = .553), which is comparable to the correlations between the confidence and accuracy factors in Tables 6 and 9. Only less than half of participants provided self-reported SAT and ACT scores. Based on these subsamples of participants, self-reported SAT and ACT scores show rather low correlations with the confidence scores (r = .271 and .238). Self-reported HS-GPA shows the lowest correlation with the confidence scores (r = .159).

Table 10

Pearson Product-Moment Correlations Between Various Accuracy Scores and TOEFL iBT Reading/Listening Confidence Scores

Variable	TOEFL iBT reading and listening confidence score
Accuracy scores	
Reading 1 & 2	.499
Listening 1 & 2	.539
Speaking	.338
Writing	.414
TOEFL total score	.552
Self-reported SAT[a]	.271
Self-reported ACT[b]	.348
HS-GPA	.159

[a] n = 384. [b] n = 342.

Correlations Between Confidence Scores and Big Five Personality Traits

Table 11 presents correlations of Big Five personality dimensions with the accuracy (i.e., total TOEFL iBT) and confidence (i.e., combined reading and listening) scores. The accuracy and confidence scores based on the TOEFL iBT exam show a moderate relationship with openness. This finding is in agreement with the work by Pallier et al. (2002). We also obtained moderate correlations of agreeableness with both the accuracy and confidence scores, which is slightly higher than what previous studies have reported (Kleitman & Stankov, 2006; Pallier et al.). In general, the pattern of personality traits' correlations is similar for the accuracy and confidence scores. What is particularly noteworthy is that contrary to previous studies showing stronger links of confidence to personality traits than any other constructs (see Blais et al., 2005; Klayman et al., 1999; Kleitman & Stankov, 2001; Pallier et al., 2002), the data show that personality traits have slightly higher correlations with the accuracy scores than with the confidence scores. This suggests that personality traits have slightly closer, relationships to cognitive abilities than they do to confidence.

Table 11

Pearson Product-Moment Correlations Between Big Five Factors and TOEFL iBT Total Accuracy Score and Reading/Listening Confidence Scores

Big Five factors	Accuracy scores	Confidence scores
	TOEFL iBT total score	Reading and listening sections
Extraversion	.037	.040
Agreeableness	.343	.234
Conscientiousness	.138	.158
Emotional stability	.046	.119
Openness	.390	.332
Average	.191	.177

Summary of Structural Findings

We have shown that reliabilities of confidence are very high, with alpha coefficients of .90s. Thus, we conclude that confidence can be measured reliably. The findings suggest that a

confidence factor is distinct and cannot be reduced to the domains of cognitive abilities, metacognition, or personality traits, although confidence is moderately correlated with them domains. In the present data, no strong evidence challenges the conclusion that confidence is indeed a separate trait.

Group Differences in Confidence: Gender, Ethnicity, and College Type

Table 12 displays the means for the accuracy, confidence, and bias scores of the TOEFL iBT reading and listening sections for the total and for male and female participants. Positive signs for the mean bias scores in Table 12 indicate that there is overconfidence at the group level for both male and female participants, especially on reading tasks. However, overall, male participants tended to show greater overconfidence bias than female participants. In this sample, the overconfidence bias for both male and female participants shows somewhat smaller magnitude than what was reported by Pallier (2003).

Table 12

Means for Accuracy, Confidence, and Bias Scores on TOEFL iBT Reading and Listening Sections, Versions 1 & 2, by Gender

Gender	*N*	Reading 1 & 2			Listening 1 & 2		
		Mean accuracy	Mean confidence	Mean bias	Mean accuracy	Mean confidence	Mean bias
Total sample	822	76.18	87.92	11.74	82.54	88.84	6.30
Male	304	74.57	87.58	13.01	79.88	88.27	8.39
Female	518	77.23	88.16	10.93	84.19	89.20	5.01
t-test[a]		2.040*	.748	2.106*	4.204**	1.482	3.930**

*$p < 0.05$. **$p < 0.01$.

[a] $df = 821$.

Table 13 presents means for ethnic groups (White, African American, and Hispanic) for the accuracy, confidence, and bias scores. The Hispanic group is in the middle on all three groups, with the White group showing the smallest bias and the African American group

showing the largest bias. This pattern is evident in both the TOEFL iBT reading and listening sections.

Table 13

Means for Accuracy, Confidence, and Bias Scores on TOEFL iBT Reading and Listening Sections, Versions 1 & 2, by Ethnicity

Ethnicity	*n*	Reading 1 & 2			Listening 1 & 2		
		Mean accuracy	Mean confidence	Mean bias	Mean accuracy	Mean confidence	Mean bias
White	605	79.32	89.02	9.69	85.27	89.52	3.92
African American	113	61.75	83.02	21.27	68.87	84.98	16.11
Hispanic	60	70.41	86.51	16.10	79.19	89.31	10.12
F-test [a]		59.41**	17.28**	21.673**	51.24**	9.30**	19.41**

**$p < 0.01$.

To nail down characteristics of people with high bias scores, we broke down the sample further by college type—those who attend 2-year versus 4-year colleges. Table 14 shows the means for the composite accuracy, confidence, and bias scores for African American students attending 2-year and 4-year colleges. While there are no significant differences on the accuracy scores, the differences on the confidence scores are significant between these two groups. African American students attending 2-year colleges show stronger bias than those attending 4-year colleges ($t[110] = 4.07$, $p < .01$).[7]

Hard-Easy Effect and Bias

One consistent pattern through Tables 12 to 14 is that the confidence scores show less variability than the accuracy scores between the groups (i.e., between the gender group, between the ethnic group, and between the college-type group). For instance, Table 13 shows that the range for the accuracy scores between the ethnic groups is 23.52, resulting from the highest score of White students on listening section (85.27) and the lowest score of African American students on reading section (61.75). On the other hand, the range for the confidence scores is only 6.50 with the highest confidence score of 89.52 on listening section from White students and the

lowest confidence score of 83.02 on the reading section from African American students. Thus, the main difference in the bias scores between the ethnic groups is due to greater differences in the accuracy scores, not due to the confidence scores.

Table 14

Means for Accuracy, Confidence, and Bias Scores on Combined Reading 1 & 2 and Listening 1 & 2 Among African American Students by College Types

College types	Combined Reading 1 & 2 and Listening 1 & 2		
	Mean accuracy	Mean confidence	Mean bias
African American students attending 2-year college	64.25	88.15	23.90
African American students attending 4-year college	66.39	81.15	14.76
t-test [a]	1.53(ns)	2.97*	4.07*

*$p < 0.05$. **$p < 0.01$.

[a] $df = 110$.

Predictive Validity of Confidence Scores

This section examines evidence for incremental predictive validity of confidence scores. In all the analyses reported here, the participants' scores for the TOEFL iBT reading and listening sections were entered as the first block of predictors, and the overall confidence scores from reading and listening, versions 1 and 2 were added in the second block. The question of interest is whether confidence accounts for additional variance in the criterion measures that remains after accuracy scores on the TOEFL iBT reading and listening section are entered into the regression equation.

The criterion measures consist of three types of cognitive outcomes: (a) scores derived from TOEFL iBT (the total TOEFL iBT score as well as speaking and writing section scores), (b) numeracy and overclaiming *d'* test scores, and (c) self-reported SAT scores and HS-GPA. Table 15 presents *R*-square statistics as a summary index for regression analyses. As can be seen in Table 15, the confidence scores derived from the TOEFL iBT reading and listening sections

contribute significantly to the incremental prediction on both the TOEFL iBT total and speaking and writing section scores. The confidence scores provide statistically significant improvement in predicting writing and speaking scores that were not used in the assessment of confidence.[8] The numeracy test also shows incremental validity for the confidence scores. The absence of evidence for incremental validity for overclaiming *d'* is not entirely surprising since the correlations between the overclaiming *d'* test and the confidence scores were comparatively low to begin with (see Table 7: $r = .254$ for the reading confidence score and $r = .168$ for the listening confidence score). That there is no evidence for incremental validity for SAT and HS-GPA may be attributable to self-reported scores by participants. In fact, information is available for only the reliability of all three criterion measures that showed incremental validity.

Table 15

Summary of Regression Analysis Results: R-Square Coefficients Showing Incremental Validity of Reading and Listening Confidence Scores in Predicting Various Accuracy Score Criteria Above and Beyond Reading and Listening Accuracy Scores

Criteria	*R*-squares from regression analysis	
	Regression model one predictors[a]	Regression model two predictors[b]
TOEFL		
Total	.875	.877[c]
Writing	.385	.395[c]
Speaking	.269	.273[c]
Numeracy	.401	.404[c]
Overclaiming	.144	.145
SAT[d]	.307	.307
HS-GPA	.079	.079

[a] Reading and listening accuracy scores. [b] Reading and listening accuracy and confidence scores. [c] Statistically significant incremental validity change from the first model to the second model; had incremental *R*-squares, indicating significant differences in the *R*-squares between the two regression models. [d] $n = 384$.

In summary, the confidence scores from the two sections of TOEFL iBT do provide incremental validity over and above their yoked accuracy scores to the three independent sets of cognitive tests scores—speaking, listening, and numeracy. We conclude that people's confidence scores can predict their cognitive abilities in some measures even after controlling for the cognitive abilities that are used as the basis for measuring confidence level. Reported incremental predictive validity is noteworthy from the theoretical point of view. Since the amount of incremental variance accounted for by the confidence is smaller than 1%, practical importance of this finding is minimal.

Discussion

Some findings in this study agree with previous literature on confidence. First, confidence measures have higher reliabilities than ability scores. Second, a confidence should be considered as a separate trait, distinct from other traits such as abilities and personality traits. The present study employed confidence scores based only on TOELF iBT, but previous work has shown that the confidence factor can be reliably extracted from a large number of diverse cognitive tests (Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997). This study also shows that confidence tends to be more closely related to cognitive abilities than it is related to personality. This may suggest that different types of confidence exist—cognitive confidence, which is captured by the procedures in this study, versus social confidence, which can be measured as a part of personality. Third, males exhibit a stronger overconfidence bias than females. Previous studies also support age-related group differences in confidence in that older people are somewhat more confident than younger people (Crawford & Stankov, 1996). Age variables were not available for this study.

Several findings in the present study extend what was reported previously. Study results indicate differential overconfidence bias among ethnic groups with African American students, particularly those attending 2-year colleges, showing more pronounced bias. Although one needs to be cautious in interpreting this finding because of the relatively small sample size in comparison to the number of White students, African American students appear to be least aware of their level of performance on tests of verbal abilities compared to the other groups in the present study.

It seems plausible that overconfidence may lead to suboptimal effort and therefore contribute to achievement gaps. A related question is the malleability of confidence. Could it be

that confidence is more responsive to intervention than cognitive abilities captured by the accuracy scores? If so, such interventions may provide means for reducing achievement gaps. The implied link is that better calibrated confidence may encourage individuals to put more effort into their study or work, which can lead to improved performance and reduced achievement gaps.

Study results also show small incremental validity of confidence scores for predicting cognitive performance on tests of writing, speaking, and numeracy. In particular, the incremental validity for the numeracy underscores that confidence can be seen as a broad, task-independent trait. However, it is important to keep in mind the fact that the amount of improvement in prediction is small and practical importance of this finding should not be overemphasized. The absence of incremental validity evidence for self-reported SAT scores and HS-GPA may be due in part to their accuracy and unknown (and possibly low) reliability of these self-reported measures.

The evidence for an incremental increase in the validity of confidence ratings suggests possible uses of confidence assessments as a way to incorporate noncognitive measures in service of selection, guidance, and intervention. The unique aspect of measuring confidence employed in this study (i.e., confidence measures attached to the answer to each test item) makes its assessment more reliable and valid in comparison to most other noncognitive measures. This yoked nature of confidence makes it difficult to fake it or to be coached to answer in a particular way. Confidence ratings and bias scores may also be used as predictors of criteria other than cognitive performance. For example, confidence may be a good predictor of dropout rates or time to completion in graduate schools since these criteria are known to be affected by factors other than cognitive abilities (see Kuncel, Hezlett, & Ones, 2001).

This study shows that the confidence scores at the group-level do not change at the same rate as the accuracy scores. For example, in Table 12 the difference between means for female and male study participants on listening accuracy is 4.31 (84.19 minus 79.88), but for listening confidence it is 0.93 (89.20 minus 88.27). Confidence does change in concert with measures of accuracy, but changes in confidence are less pronounced than changes in accuracy. It seems that individuals have tendency to fail to judge the degree of their inability to solve test problems. Thus, overconfidence bias results from relative consistency of the confidence scores.

What is known as a *hard-easy effect* in decision-making literature (see Suantak, Bolger, & Ferrell, 1996) may be one possible explanation for the bias scores. Put simply, the magnitude of overconfidence bias depends on task difficulty, and overconfidence bias tends to be more pronounced as the task becomes more difficult. In this paper, the hard-easy effect is obvious not only from the group comparisons but also from task comparisons. For example, arithmetic means for the whole sample (displayed in Table 1) indicate that the reading section is more difficult than the listening section but the average confidence differs little between the listening and reading sections. As a result, the reading section shows an overconfidence bias that is more than twice as large as in the listening section.

Although the present data support this hard-easy effect interpretation, there is other empirical evidence that points to the differences in the *nature of the task* as being an important factor in the overconfidence bias (see Juslin & Olsson, 1997; Olsson & Winman, 1996). This evidence suggests that some tasks, such as visual sensory acuity measures, are insensitive to hard-easy manipulations and thus do not exhibit overconfidence bias. Because humans are wired to process information differently depending on whether the tasks are cognitive or sensory in nature, people tend to show more overconfidence in cognitive than on visual sensory tasks (Juslin & Olsson; Olsson & Winman). This suggests that the overconfidence bias observed in this study may may be due in part to sensory-cognitive difference, not just the hard-easy effect.

There are still quite a few unanswered questions about confidence. For example, can general feedback on test performance reduce excessive confidence? How would positive or negative feedback on an item affect confidence on the next item? Given that confidence has an incremental validity for predicting scores on the TOEFL iBT speaking and writing sections, does confidence have a particular value in language learning? How about personality measures other than the Big Five? Is a personality trait expressed in a social setting related to cognitive confidence?

For over a century, the study of individual differences has focused on uncovering the dimensions that can be used to understand the psychological makeup of the human species. Thus far, personality and ability domains have been mapped out reasonably well (see Carroll, 1993; Saucier & Goldberg, 2002). Every candidate for a new dimension in individual differences should be compared to what is known so far and a robust proof of its convergent and discriminant validity is the central issue. The findings reported in this paper, together with the

findings from previous studies, indicate that confidence is indeed a psychological trait that is related to, but distinct from both personality and ability traits. Within the structure of all other individual differences dimensions, confidence should be located in-between these two domains.

References

Angelis, J. C., Swinton, S. S., & Cowell, W. R. (1979). *The performance of non-native speakers of English on TOEFL and verbal aptitude tests* (TOEFL Report No. RR-03). Princeton, NJ: ETS.

Angoff, W. H., & Sharon, A. T. (1971). A comparison of scores earned on the Test of English as a Foreign Language by Native American college students and foreign applicants to U.S. colleges. *TESOL Quarterly, 5,* 129–136.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception and Psychophysics, 61,* 1369–1383.

Blais, A.-R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative tasks: The role of cognitive styles. *Personality and Individual Differences, 38,* 1701–1713.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

Crawford, J., & Stankov, L. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21,* 971–986.

Festinger, L. (1943a). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32,* 291–306.

Festinger, L. (1943b). Studies in decision: II. An empirical test of a quantitative theory of decision. *Journal of Experimental Psychology, 32,* 411–432.

Fullerton, G. S., & Cattell, J. M. (1892). On the perception of small differences (*University of Pennsylvania Philosophy Series No. 2).* Philadelphia: University of Pennsylvania Press.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences, 1,* 78–82.

International personality item pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. (n.d.). Retrieved June 28, 2005, from http://ipip.ori.org/

Johnson, D. C. (1977). The TOEFL and domestic students: conclusively inappropriate. *TESOL Quarterly, 11,* 79–86.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 10,* 344–366.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Kirsch, I., Yamamoto, K., Norris, N., Rock, D., Jungleblut, A., & O'Reilly, P. (2001). *Technical report and data file user's manual for the 1992 National Adult Literacy Survey* (NCES-2001-457). Washington, DC: National Center for Education Statistics.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior & Human Decision Processes, 79,* 216–247.

Kleitman, S. (2003). Self-confidence and self-monitoring aspects of metacognition, their nature and their place within rationality debates: An individual differences perspective. Unpublished doctoral dissertation, University of Sydney, Australia.

Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology, 15,* 321–341.

Kleitman, S., & Stankov, L. (2006). *Self-confidence and metacognitive processes.* Manuscript submitted for publication.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Lichtenstein, S., & Fischoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Lichtenstein, S., Fischoff. B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Ed.), *Judgments under uncertainty: Heuristics and biases.* Hillsdale, NJ: Erlbaum.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*(1), 129–149.

Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist, 20*(3), 107–123.

Olsson, H., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Perception and Psychophysics, 58,* 374–383.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex roles, 48,* 265–276.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). Individual differences in the realism of confidence judgments. *Journal of General Psychology, 129,* 257 –300.

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence, 32,* 297–314.

Saucier, G., & Goldberg, L. R. (2002). Assessing the big five: Applications of 10 psychometric criteria to the development of marker scales. In B. de Raad & M. Perugini (Eds.), *Big Five assessment* (pp. 29–58). Ashland, OH: Hogrefe & Huber.

Sawaki, Y., Stricker, L., & Oranje, A. (in press). *Factor structure of the TOEFL Internet-based test (iBT).* Princeton, NJ: ETS.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19,* 460–475.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117–137.

Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences, 10,* 29–50.

Stankov, L. (1999). Mining on the “no man’s land” between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 315–337). Washington, DC: American Psychological Association.

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28,* 121–143.

Stankov, L., (in press). Dimensions of cultural differences: personality, social attitudes, values and social norms. *Journal of Individual Differences.*

Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21*, 971–986.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*, 93–109.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. London: Lawrence Erlbaum Associates.

Stricker, L. J. (2002). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing, 21,* 146–173.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67,* 201–221.

Tobias, S., & Everson, H. T. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147–22). Lincoln, NE: Buros Institute of Mental Measurements.

Tobias, S., & Everson, H. T. (2002). *Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring* (College Board Research Rep. No. 2001-03). New York: College Board.

Trow, W. C. (1923). The psychology of confidence. *Archives of Psychology, 67*, 47–71.

Vickers, D. (1979). *Decision processes in visual perception.* New York: Academic Press.

Notes

[1] In our previous work, the term *self-confidence* is used in the same way as *confidence* in this paper.

[2] Stankov & Crawford (1996) mentioned that another label for the bias score, that is, realism (of confidence ratings), is sometimes employed.

[3] This is just an example. We do not expect to predict HS-GPA particularly well by aspects of language proficiency captured by the TOEFL iBT.

[4] The total *N* for this study was, in fact, 950. Preliminary screening of the data showed that a proportion of participants produced patterns of responses indicating careless responding (e.g., the same answer was provided for all items in a given instrument). This tendency was particularly pronounced with the instruments administered in the afternoon session, not with TOEFL iBT scores. Our criterion for excluding participants was the presence of evidence for careless responding in 3 out of 28 instruments administered in the afternoon session, some of which are used in the present report.

[5] One of the items in the reading, version 2 was based on a partial scoring method (i.e., instead of a 0,1 scoring key, a 0, 1, 2, 4 scoring key was employed). When we removed this item from the analyses, the overall bias score for reading, version 2 was reduced by 2.94%, closer to the borderline value for noteworthy bias scores.

[6] The correlation between listening and reading absolute change scores is somewhat high (r = .60), indicating the tendency for participants' scores to change between the two testing sessions is consistent across the two sections.

[7] For White and Hispanic students, the difference between 2- and 4-year colleges on bias was not statistically significant.

[8] In the regression analyses of this paper, we also entered an interaction score (i.e., a multiple of accuracy and confidence scores). This interaction score did not contribute significantly to any of the criterion measures (rows) in Table 15.